CANDID: Comparison Algorithm for Navigating Digital Image Databases

P. M. Kelly and T. M. Cannon
Computer Research and Applications Group, CIC-3
Los Alamos National Laboratory
Los Alamos, NM, 87545

Abstract

In this paper, we propose a method for calculating the similarity between two digital images. A global signature describing the texture, shape, or color content is first computed for every image stored in a database, and a normalized distance between probability density functions of feature vectors is used to match signatures. This method can be used to retrieve images from a database that are similar to an example target image. This algorithm is applied to the problem of search and retrieval for a database containing pulmonary CT imagery, and experimental results are provided.

1 Introduction

Future data management systems will be required to handle not only textual data, but also massive amounts of non-textual data such as raw system measurements, digital imagery, sound samples, and video clips. These systems will be extremely valuable if they can provide easy access to this diversity of data. Unfortunately, many systems will be simple archives where diverse types of data can only be retrieved by searching for desired dates, titles, subject keywords, and associated textual descriptions. The value of these systems can be greatly enhanced by adding the ability to search directly on the non-textual data, instead of searching only on the associated textual metadata.

Content-based retrieval of digital imagery is currently an active area of research. Several methods have been proposed for the comparison of pictorial or iconic images [1, 10]. Other methods compare the relative geometries and positions of different objects in each image [4, 8]. In the QBIC Project [3, 9], color, texture, and shape features are computed for each "object" in an image, as well as for each image. A Euclidean distance measure is then used to determine similarity between objects or images. We have developed a new method for comparing digital images that is based on ideas being explored for searching databases containing free-text documents.

Modern databases typically use keywords to search through large amounts of textual data. Although these techniques work well, a user is required to fully understand what is being sought by providing specific keywords on which to search. Some newer methods for searching textual databases use "global signatures" to represent the content (or topic) of an entire document instead of using a keyword indexing scheme. An example is the N-gram approach to document fingerprinting [11].

When using the N-gram method for document comparison, a global signature is computed for each document in the database. This signature represents the content, or topic, of a document in an abstract sense. A signature is typically represented by a histogram of the number of times that each substring of length N occurs in the document, where N is a predetermined value. As an example, for a case-insensitive alphabet of 26 letters, there are 26^3, or 17,576, different trigrams ("aaa", "aab", "aac", $\cdots$, "zzz"). The signature for each document in this example is therefore a normalized vector of dimension 17,576. A dot-product between N-gram signatures determines the similarity between any two documents. Using this approach for retrieving documents from a database, a user can pose queries such as, "Show me all of the documents that are similar to this example". A user does not need to identify which specific keywords or phrases should be searched on.

We are finding that this technique of using a global signature to characterize an entire set of data is also very useful in retrieving non-textual data such as digital imagery. The *CANDID* algorithm (Comparison Algorithm for Navigating Digital Image Databases) presented in this paper is analogous to the N-gram approach described above in the sense that we attempt

to describe an entire image with a global signature, and then match signatures with some distance measure to determine image similarity. Each image stored in the database is characterized by a global signature that can represent features such as textures, shapes, and colors. When the database retrieval software is asked to search for images similar to a given target image, it first computes the global signature of the target image and then matches it against the signatures of all images in the database. A handful of images having similar content, i.e. database images having a similar signature to the target image, is returned to the user. A normalized distance between probability density functions of feature vectors is used to match signatures.

2 Signature Computation

We must first recognize that similarity between images is an abstract concept; making judgements is very subjective. As an example, consider three different color pictures in an automobile magazine. One reader may think that image A is more similar to image B than to image C because both A and B contain red automobiles, and C contains a blue automobile. A second reader, however, might claim that image C resembles image A more than image B does, because the cars in both A and C are convertibles, whereas the car in B is not.

With this in mind, it is important to approach the problem of image comparison differently for every application. Shape descriptors, color features, and texture measures are all able to represent some of the information contained in an image, but the way in which they are used determine what we mean when we say that two images are "similar". The feature selection process for any application is one of the most important aspects in solving the problem.

In contrast to the QBIC method [3, 9] where color, shape, and texture measurements are calculated for the entire image or for each user-specified object, we take an approach that more closely resembles the N-gram work for textual data. The general idea is that we first compute several features (local color, texture, and/or shape) at every pixel in the image, and then make a "histogram" of feature vector (pixel vector) occurrences for that image. Unlike textual data, we will most likely not have a finite number of unique feature vectors that can occur in our data, and therefore calculate a continuous probability density function over the multidimensional feature space instead of an actual histogram.

Probability density function estimation is a large problem in itself; we attempt to estimate the probability density function as a weighted sum of gaussians. A gaussian distribution function is defined by a mean vector $\underline{\mu}_i$ and a covariance matrix Σ_i. A general data clustering routine can provide clusters for which for $\underline{\mu}_i$ and Σ_i can be obtained. We use the method proposed in [5] which suggests using the k-means clustering algorithm [2, 12] followed by an optional cluster merging process. A mean vector and covariance matrix are computed for each of the resultant clusters, and the associated gaussian distribution function is weighted by the number of elements in the corresponding cluster.

3 Signature Comparison

Consider the problem of comparing two images I_1 and I_2 for which we have already calculated probability density functions $P_{I_1}(\underline{x})$ and $P_{I_2}(\underline{x})$ that describe the distribution of feature vectors over the N-dimensional vector space for I_1 and I_2, respectively. We can then define a distance measure to compare these distributions:

$$dist\,(I_1, I_2) \;=\; \left[\int_{\Re} \left(\, P_{I_1}(\underline{x}) - P_{I_2}(\underline{x}) \,\right)^2 d\underline{x}\right]^{\frac{1}{2}}$$

This is an infinite integral over the entire N-dimensional feature space. If we assume that we have estimated these probability density functions by a weighted sum of gaussians as proposed above, each distribution takes the form:

$$\begin{aligned} P_I(\underline{x}) \;&\approx\; \sum_{i=1}^{K} w_i G(\underline{x}, \underline{\mu}_i, \Sigma_i) \\ &=\; \sum_{i=1}^{K} w_i G_i(\underline{x}) \end{aligned}$$

We will differentiate between P_{I_1} and P_{I_2} with the following notation:

$$P_{I_1} = \sum_{i=1}^{K_1} w_i G_i(\underline{x}) \qquad P_{I_2} = \sum_{i=1}^{K_2} v_i F_i(\underline{x})$$

Substituting this probability density function estimate into our distance measure, we get:

$$dist\,(I_1, I_2) = \left[\int_{\Re} \left(\sum_{i=1}^{K_1} w_i G_i(\underline{x}) \;-\; \sum_{i=1}^{K_2} v_i F_i(\underline{x})\right)^2 d\underline{x}\right]^{\frac{1}{2}}$$

We can expand this equation to see exactly how to calculate it:

$$
\begin{aligned}
dist\,(I_1, I_2) \;=\; & \Bigg[\sum_{i=1}^{K_1} w_i^2 \int_{\Re} G_i^2(\underline{x})\, d\underline{x} + \\
& \sum_{i=1}^{K_2} v_i^2 \int_{\Re} F_i^2(\underline{x})\, d\underline{x} + \\
& 2 \sum_{i=1}^{K_1} \sum_{j=i+1}^{K_1} w_i w_j \int_{\Re} G_i(\underline{x}) G_j(\underline{x})\, d\underline{x} + \\
& 2 \sum_{i=1}^{K_2} \sum_{j=i+1}^{K_2} v_i v_j \int_{\Re} F_i(\underline{x}) F_j(\underline{x})\, d\underline{x} - \\
& 2 \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} w_i v_j \int_{\Re} G_i(\underline{x}) F_j(\underline{x})\, d\underline{x} \Bigg]^{\frac{1}{2}}
\end{aligned}
$$

This distance measure contains $O(K_1^2 + K_2^2)$ terms, where each term contains an infinite integral over the product of two gaussians. It can be shown that:

$$
\begin{aligned}
\int_{\Re} G_i(\underline{x})\, G_j(\underline{x})\, d\underline{x} \;=\; & (2\pi)^{-\frac{N}{2}} \left| \Sigma_i + \Sigma_j \right|^{-\frac{1}{2}} \cdot \\
& \exp\left[-\frac{1}{2}(c_1 + c_2) \right]
\end{aligned}
$$

where c_1 and c_2 are given by:

$$
\begin{aligned}
c_1 \;=\; & \underline{\mu_i}^T \Sigma_i^{-1} \underline{\mu_i} \;+\; \underline{\mu_j}^T \Sigma_j^{-1} \underline{\mu_j} \\
c_2 \;=\; & -(\underline{\mu_i}^T \Sigma_i^{-1} + \underline{\mu_j}^T \Sigma_j^{-1}) \cdot \\
& \left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)^{-1} \cdot \\
& (\underline{\mu_i}^T \Sigma_i^{-1} + \underline{\mu_j}^T \Sigma_j^{-1})^T
\end{aligned}
$$

Furthermore, for the special case where $\underline{\mu_i} = \underline{\mu_j}$ and $\Sigma_i = \Sigma_j$, we can simplify this even further:

$$
\int_{\Re} G_i^2(\underline{x}) d\underline{x} \;=\; 2^{-N} \pi^{-\frac{N}{2}} \left| \Sigma_i \right|^{-\frac{1}{2}}
$$

In order to interpret $dist\,(I_1, I_2)$, we normalize it between 0 and 1. This can be achieved by noting that this distance is maximized when there is no overlap between the two probability density functions. We divide $dist\,(I_1, I_2)$ by the following value, which we will call $max\,(I_1, I_2)$:

$$
\begin{aligned}
max\,(I_1, I_2) \;=\; & \Bigg[\sum_{i=1}^{K_1} w_i^2 \int_{\Re} G_i^2(\underline{x})\, d\underline{x} + \\
& \sum_{i=1}^{K_2} v_i^2 \int_{\Re} F_i^2(\underline{x})\, d\underline{x} + \\
& 2 \sum_{i=1}^{K_1} \sum_{j=i+1}^{K_1} w_i w_j \int_{\Re} G_i(\underline{x}) G_j(\underline{x})\, d\underline{x} + \\
& 2 \sum_{i=1}^{K_2} \sum_{j=i+1}^{K_2} v_i v_j \int_{\Re} F_i(\underline{x}) F_j(\underline{x})\, d\underline{x} \Bigg]^{\frac{1}{2}}
\end{aligned}
$$

4 Experimental Results

We have used this concept of global signature matching to retrieve medical imagery based on image content. Pulmonary CT scans reveal the gross pathology indicative of diseased lung tissue resulting from a variety of disorders such as lymphangioleiomyomatosis (LAM), idiopathic pulmonary fibrosis (IPF), scleroderma, asthma, and vasculitis. Since CT data is acquired digitally, it can be easily stored in a computer database. It would be a natural extension of this process to search a database to retrieve images that exhibit the same pathology as the current study. These images would provide the radiologist with immediate access to past cases where similar problems were encountered, thereby aiding with the current patient's diagnosis and treatment.

We applied *CANDID* to this problem of retrieving pulmonary CT imagery from a database containing a total of 152 lung images taken from pulmonary CT studies of 9 different patients (see Table 1). Each image was 512×512 pixels in size, consisting of 16-bit data, and was reconstructed from the original tomographic projections at either standard resolution, in which the entire chest cavity is visible, or at a retargeted resolution, where the focus of the image was on a single lung. As shown in Table 1, about 41% of the images in our database were reconstructed at a retargeted resolution.

For this application, we are primarily interested in retrieving images containing similar textures. The features selected for this problem were texture energy measures developed by Laws [6, 7], which have the advantage of being able to discriminate between different textures, while being quick and easy to compute. We generated a full set of texture features, and then attempted to find a small subset of these features that

would allow *CANDID* to successfully match images of lungs affected by the same disease.

Patient	Diagnosis	Num Images	Resolution
1	LAM	37	Standard
2	LAM	6	Retargeted
3	LAM	30	Retargeted
4	IPF	24	Retargeted
5	Normal	2	Retargeted
6	Scleroderma	10	Standard
7	Vasculitis	27	Standard
8	Asthma	8	Standard
9	Asthma	8	Standard

Table 1: Contents of CT Image Database

To generate a global texture signature describing an image, we first calculated texture features for each pixel, which was done in three steps. In step one, we convolved the image with a number of Laws' convolution kernels, then each pixel value was replaced by the sum of the absolute values of the pixel values in a square neighborhood surrounding it:

$$I_{new}(x,y) = \sum_{i=x-N}^{x+N} \sum_{j=y-N}^{y+N} \mid I_{old}(i,j) \mid$$

Finally, related features were added together to provide features invariant to rotation.

L5	=	[	1	4	6	4	1]
E5	=	[	-1	-2	0	2	1]
S5	=	[	-1	0	2	0	-1]
W5	=	[	-1	2	0	-2	1]
R5	=	[	1	-4	6	-4	1]

Table 2: Center-Weighted Vectors

Table 2 lists the 5 one-dimensional center-weighted convolution kernels which are used to create the 25 two-dimensional 5-by-5 convolution kernels. The names of these one-dimensional kernels are mnemonics for Level, Edge, Spot, Wave, and Ripple. Each two-dimensional kernel is created by convolving a horizontal kernel with a vertical kernel. For instance, an E5L5 kernel is formed by convolving a horizontal E5 kernel with a vertical L5 kernel.

After convolving the original image with one of the 5-by-5 convolution kernels, the associated Texture Energy Measure (TEM) for each pixel is calculated by summing the absolute pixel values of the convolved image within a 15x15 pixel window. A total of 25 TEM images were calculated during this stage of image processing. This set was then reduced by combining related TEM images, such as the L5E5 and E5L5 images, the S5R5 and R5S5 images, etc. The E5E5, S5S5, W5W5, and R5R5 TEM images, which were not combined with other images, were scaled by a factor of two in order to keep them consistent with the other TEM images. All images were divided by the L5L5 image to normalize features for contrast, as suggested in [7], after which the L5L5 image was discarded. The result was a set of 14 images, each representing some texture feature for the image. Each pixel in the image is now represented by a vector of 14 features.

From these features, we computed three different sets of signatures. Set A was generated using the following four TEM features: E5E5, S5S5, W5W5, and R5R5. Set B used the four features from set A along with the L5E5, L5S5, L5W5, and L5R5 features, for a total of eight features. Finally, set C was generated using all 14 texture measures. All signatures consisted of a weighted sum of 20 gaussian distributions, found by using the k-means clustering algorithm to cluster the feature vectors in each image.

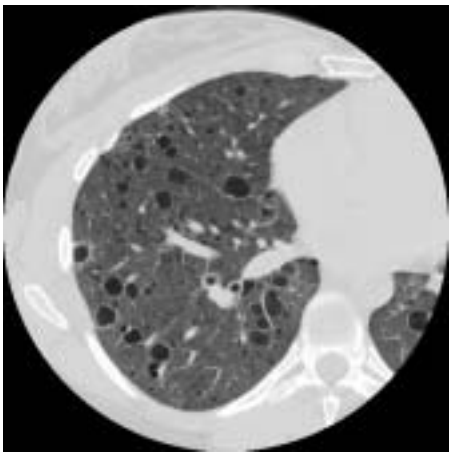

Figure 1: First Target Image: Patient 2, CT Image 0. This patient has LAM disease, which causes the formation of empty cysts in the lung.

The first example target image is shown in Figure 1. This is an image of a lung that has been affected by LAM. We queried our database using *CANDID*, and retrieved the 15 best matches to the target image from the database. We did this for each of our three signature sets A, B, and C. For each case, the best 15 matches to the target image were high-resolution images from patients 2 and 3, who were also diagnosed

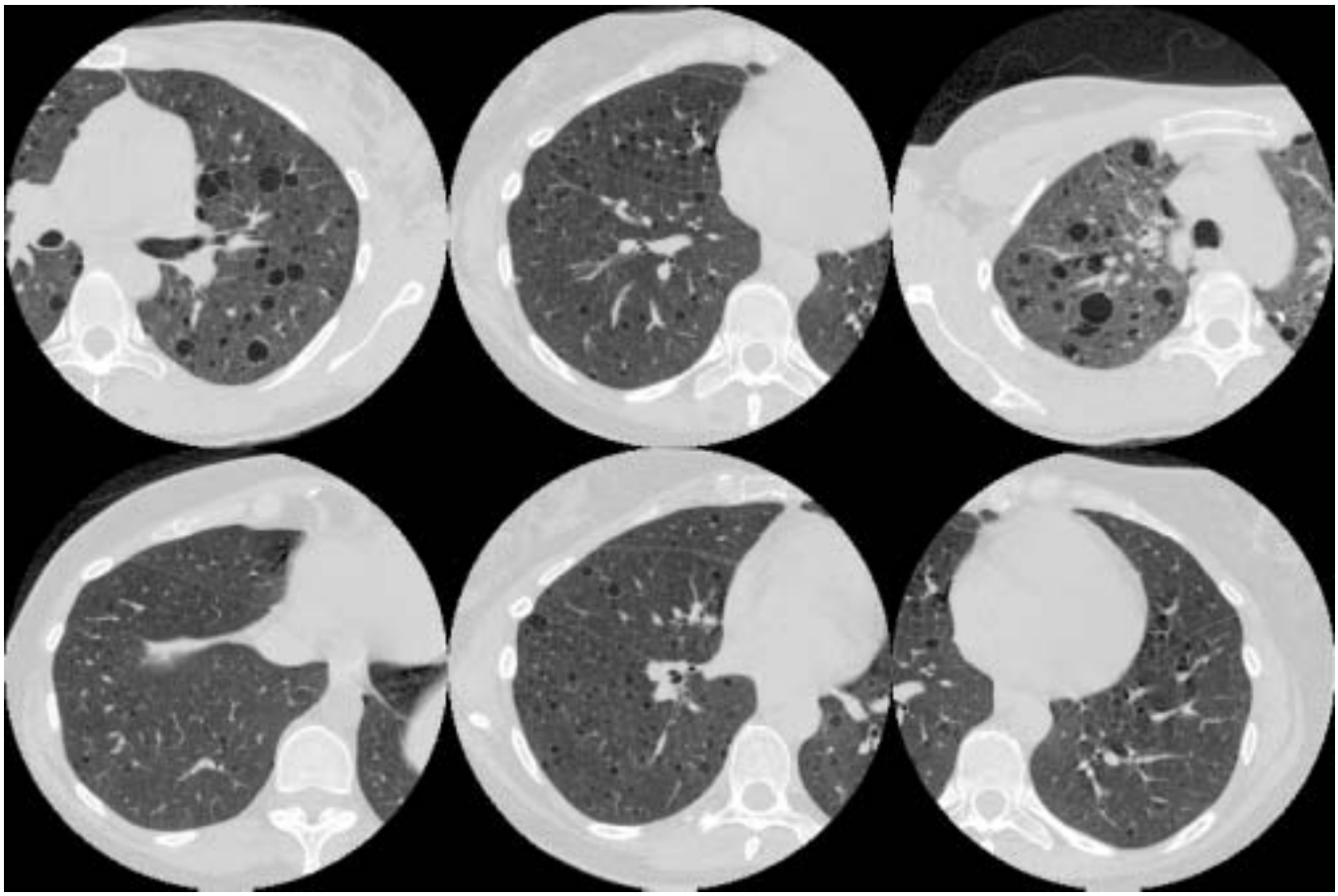

Figure 2: The best 6 matches to the first target image were also images of lungs afflicted with LAM disease.

as having LAM. The best 6 matches when using signature set A are displayed in Figure 2. These results using signature set A were as good as the results using sets B and C.

Figure 3 shows the second target image, which shows a lung affected by IPF (patient 4). As before, we employed *CANDID* to retrieve the 15 most similar images from the database. Again, for each of our three signature sets A, B, and C, the best 15 matches were all images taken from patient 4 (diagnosed as having IPF). The best 6 matches using signature set A for this example are displayed in Figure 4. Again, the results using signature set A were as good as the results using sets B and C.

We ran *CANDID* on a Sun SPARCstation IPX. To compare a single target image to all 152 database images took 32 seconds of CPU time when using signature set A (4-dimensional data), 71 seconds when using signature set B (8-dimensional data), and 163 seconds when using signature set C (14-dimensional data).

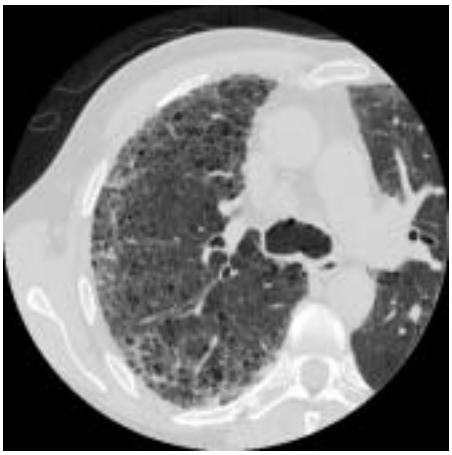

Figure 3: Second Target Image: Patient 4, CT Image 5. This patient suffers from idiopathic pulmonary fibrosis (IPF).

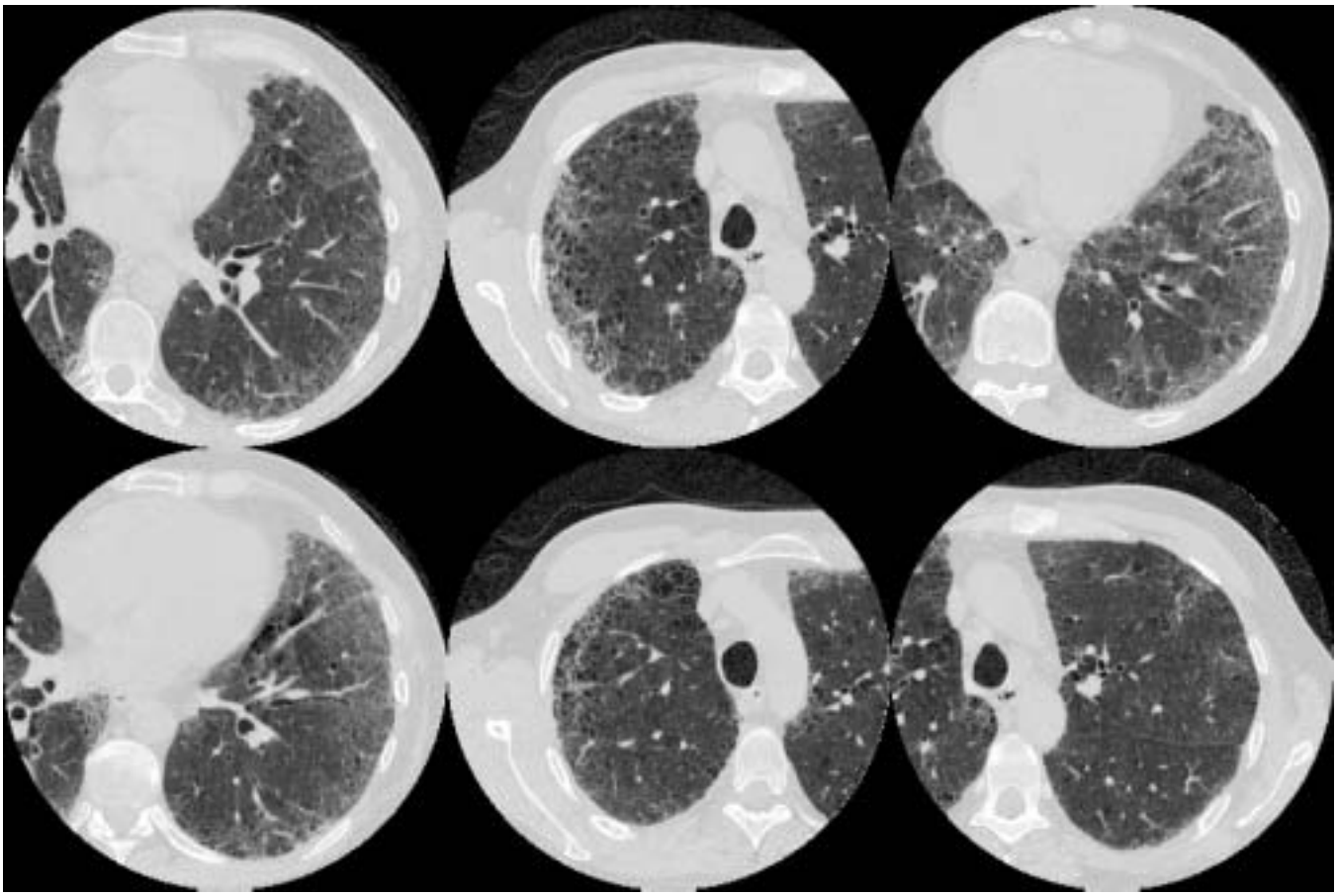

Figure 4: The best 6 matches to the second target image were also images of lungs afflicted with IPF.

It is important to note that for these experiments, the signatures contain texture information about each *entire* CT image. It considers textures around the ribs, spinal cord, and heart as well as the textures inside of each lung. This results in a system that attempts to retrieve images containing lungs of the same size and at the same resolution.

5 Conclusions

The problem of retrieving digital images from a database based on image content can be a difficult one. Not only must the meaning of "similarity" between images be determined for each application, but an algorithm must be developed to retrieve images in a manner consistent with the way that a human operator would. The *CANDID* algorithm performed extremely well on our sample database of 152 pulmonary CT images. Using only four texture features, the system successfully discriminated between different pulmonary diseases, returning images with the same content and resolution as the target images.

The general approach described in this paper is not limited to image retrieval problems. Since it attempts to characterize the distribution of features vectors in an abstract feature space, this approach can be used to work with almost any type of data and features. As an example, *CANDID* might be applied to the problem of 1-D signal matching. Many features (such as local frequency) can be computed at different positions along each signal. A signature for each signal could then be calculated and manipulated in a manner consistent with the approach we have presented.

Acknowledgements

We would like to thank Dr. John Newell and Dr. David Lynch at the National Jewish Center for Immunology and Respiratory Medicine for providing us with data. This work was performed under a U.S. Government contract (W-7405-ENG-36) by Los

Alamos National Laboratory, which is operated by the University of California for the U.S. Department of Energy.

References

[1] C.C. Chang and T.C. Wu. Retrieving the most similar symbolic pictures from pictorial databases. *Information Processing and Management*, 28(5):581–588, 1992.

[2] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, NY, 1973.

[3] C. Faloutsos, M. Flickner, W. Niblack, D. Petkovic, W. Equitz, and R. Barber. Efficient and effective querying by image content. Technical Report RJ 9453 (83074), IBM Almaden Research Center, San Jose, CA, 1993.

[4] T.Y. Hou, A. Hsu, P. Liu, and M.Y. Chiu. A content-based indexing technique using relative gometry features. In *SPIE Vol. 1662 Image Storage and Retrieval*, pages 607–720, 1992.

[5] P.M. Kelly, D.R. Hush, and J.M. White. An adaptive algorithm for modifying hyperellipsoidal decision surfaces. *Journal of Artificial Neural Networks*. In Press.

[6] K. Laws. Rapid texture identification. In *SPIE Vol. 238 Image Processing for Missile Guidance*, pages 376–380, 1980.

[7] K. Laws. *Textured Image Segmentation*. Ph.D. dissertation, Univ. of Southern Calif., January 1980.

[8] S.Y. Lee and F.J. Hsu. Spatial reasoning and similarity retrieval of images using 2D c-string knowledge representation. *Pattern Recognition*, 25(3):305–318, March 1992.

[9] W. Niblack, R. Barder, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Yaubin. The QBIC project: Querying images by content using color, texture, and shape. In *SPIE Vol. 1908 Storage and Retrieval for Image and Video Databases*, pages 173–181, 1993.

[10] F. Rabitti and P. Savino. Automatic image indexation to support content-based retrieval. *Information Processing and Management*, 28(5):547–565, 1992.

[11] T.R. Thomas. Document retrieval from a large dataset of free-text descriptions of physician-patient encounters via N-gram analysis. Technical Report LA-UR-93-0020, Los Alamos National Laboratory, Los Alamos, NM, 1993.

[12] J.T. Tou and R.C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley, Reading, MA, 1974.